Instance-Based Prediction with Guaranteed Confidence

Eyke Hüllermeier[1]

Abstract. Instance-based learning (IBL) algorithms have proved to be successful in many applications. However, as opposed to standard statistical methods, a prediction in IBL is usually given without characterizing its confidence. In this paper, we propose an IBL method that allows for deriving set-valued predictions that cover the correct answer (label) with high probability. Our method makes use of a formal model of the heuristic inference principle suggesting that similar instances do have similar labels. The focus of this paper is on the prediction of numeric values (regression), even though the method is also useful for classification problems if a reasonable similarity measure can be defined on the set of classes.

1 INTRODUCTION

As opposed to inductive, model-based machine learning methods, instance-based learning (IBL) [3, 1] provides a simple means for realizing *transductive* inference [14], that is inference "from specific to specific": Rather than inducing a general model (theory) from the data, the data itself is simply stored [13]. The processing of the data is deferred until a prediction is actually requested (or some other type of query must be answered), a property that qualifies IBL as a *lazy* learning method [2]. Predictions are then derived by combining the information provided by the stored examples in one way or other.

Typically, IBL is applied to classification problems, where predictions are derived from the query's k nearest neighbors through majority voting. Still, by combining the neighbors' predictions using a weighted sum rather than majority voting, IBL can also be employed for the estimation of numeric values [4]. In [9], the predictive performance of (numeric) IBL was found to be quite able to compete against linear regression (LR). More importantly, the authors correctly emphasized a key advantage of IBL, namely the fact that it does not assume strong (structural) properties of the data-generating process, such as linearity in LR.

This advantage, however, does not come for free. For methods that dispose of an underlying (statistical) model it is usually much simpler to quantify the *credibility* of a prediction. In LR, for example, an estimated model can be used for deriving a *confidence interval* covering a predicted output with a certain probability. Roughly speaking, this becomes possible by transferring the credibility of the model itself, estimated on the basis of the data in conjunction with the model assumptions, to predictions thereof.

In this paper, we propose an extension of IBL that allows for deriving "credible" predictions, thereby combining advantages from both instance-based and model-based learning. This extension draws its inspiration from statistical methods: The basic idea is to derive a *credible set* of predictions, which is likely to contain the correct answer, rather than making a single prediction (point-estimation). Note that this approach to credible estimation is different from characterizing the reliability of a single (point) estimation [10, 11].

The remainder of the paper is organized as follows: After some preliminaries, we introduce two concepts called *similarity profile* and *similarity hypothesis* (Section 3). On the basis of these concepts, an instance-based learning method for deriving credible sets of predictions is developed. Section 4 presents a probabilistic extension of the method, which improves its performance and robustness in real-world applications. Finally, some experimental results are discussed in Section 5. The work presented here is an extension and continuation of [7, 8].

[1] Department of Mathematics and Computer Science, Marburg University, Germany, email: eyke@informatik.uni-marburg.de

2 PRELIMINARIES

Let $\mathcal{X}$ denote an instance space, where an instance corresponds to the description x of an object (usually in attribute–value form). We assume $\mathcal{X}$ to be endowed with a reflexive and symmetric similarity measure $\sigma_{\mathcal{X}}$. $\mathcal{L}$ is a set of labels, also endowed with a reflexive and symmetric similarity measure, $\sigma_{\mathcal{L}}$. Both measures shall be normalized such that $0 \leq \sigma_{\mathcal{X}}, \sigma_{\mathcal{L}} \leq 1$. $\mathcal{D}$ denotes a sample (memory, case base) that consists of n labeled instances (cases) $\langle x_\imath, \lambda_{x_\imath} \rangle \in \mathcal{X} \times \mathcal{L}$, $1 \leq \imath \leq n$. Finally, a novel instance $x_0 \in \mathcal{X}$ (a query) is given, whose label λ_{x_0} is to be predicted.

We do not make any assumptions on the cardinality of the label set $\mathcal{L}$. In fact, we do not even distinguish between the performance tasks of classification (estimating one among a finite set of class labels) and regression (estimating a real-valued output), which means that $\mathcal{L}$ might even be infinite. Subsequently, we therefore employ the term *label* as a generic term not only for the name of a class in classification but also for numeric values in regression.

3 CREDIBLE INSTANCE-BASED LEARNING

3.1 Similarity Profiles

A key idea of our approach is to proceed from a formal model of the heuristic IBL assumption, suggesting that similar instances do have similar labels. This formalization will provide the basis of a sound inference procedure including assertions about the confidence of predictions. To begin, suppose that the IBL assumption has the following concrete meaning:

$$\forall\, x, y \in \mathcal{X} \,:\, \zeta\left(\,\sigma_{\mathcal{X}}(x,y)\,\right) \,\leq \sigma_{\mathcal{L}}\left(\lambda_x, \lambda_y\right), \qquad (1)$$

where ζ is a function $[0,1] \rightarrow [0,1]$. This function assigns to each similarity degree between two instances, α, the largest similarity degree $\beta = \zeta(\alpha)$ such that the following property holds: The labels of two α-similar instances are guaranteed to be at least β-similar. We call ζ the *similarity profile* of the application at hand. More formally,

ζ is defined as follows: For all $\alpha \in [0,1]$,

$$\zeta(\alpha) =_{\text{def}} \inf_{x,y \in \mathcal{X}, \sigma_{\mathcal{X}}(x,y)=\alpha} \sigma_{\mathcal{L}}\left(\lambda_x, \lambda_y\right). \tag{2}$$

Note that the similarity profile conveys a precise idea of the extent to which an application actually meets the IBL assumption. Roughly speaking, the "larger" ζ is, the better this assumption is satisfied.

The constraint (1) suggests the following inference scheme for predicting the label λ_{x_0}:

$$\lambda_{x_0} \in C(x_0) =_{\text{def}} \bigcap_{\imath=1}^{n} \mathcal{N}_{\zeta(\sigma_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath}), \tag{3}$$

where the β-neighborhood $\mathcal{N}_\beta(\lambda)$ of a label λ is given by the set $\{\lambda' \in \mathcal{L} \,|\, \sigma_{\mathcal{L}}(\lambda,\lambda') \geq \beta\}$. This inference scheme is obviously correct in the sense that $C(x_0)$ is guaranteed to cover λ_{x_0}, a property that follows immediately from the definition of the similarity profile ζ. We call $C(x_0)$ a *credible label set* and refer to the inference scheme itself as CIBL (Credible IBL).

A similarity profile can also be attached to single cases: The *local similarity profile* $\zeta_\imath$ of the $\imath-$th case $\langle x_\imath, \lambda_{x_\imath}\rangle$ is defined as in (2), except that the infimum is taken only over those pairs of cases that involve the $\imath-$th case itself. Thus, a local profile indicates the validity of the IBL assumption for *individual* cases (and might hence serve as a criterion for selecting "competent" cases to be stored in the memory $\mathcal{D}$ [12]). In the inference scheme (3), the neighborhoods $\mathcal{N}_{\zeta(\sigma_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath})$ are replaced by the neighborhoods $\mathcal{N}_{\zeta_\imath(\sigma_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath})$. We refer to this type of local inference as CIBL-L.

Mathematically speaking, a profile ζ is the lower envelope of all individual profiles $\zeta_\imath$. Consequently, CIBL-L will usually yield predictions that are more precise than those of CIBL. The price to pay is a higher computational complexity, since a profile must be maintained for every case in the memory $\mathcal{D}$.

3.2 Similarity Hypotheses

The application of the inference scheme (3) requires the similarity profile ζ (resp. the local profiles $\zeta_\imath$) to be known, a requirement that will usually not be fulfilled. This motivates the related concept of a (local) *similarity hypothesis*, a function $h : [0,1] \to [0,1]$, which is thought of as an approximation of a similarity profile. A hypothesis h is called *stronger* than a hypothesis h' if $h' \leq h$ and $h \not\leq h'$. We say that h is *admissible* if $h(\alpha) \leq \zeta(\alpha)$ for all $\alpha \in [0,1]$.

It is obvious that using an admissible hypothesis h in place of the true similarity profile ζ within the inference scheme (3) leads to correct predictions $C^{est}(x_0) \supseteq C(x_0)$. Indeed, $h \leq \zeta$ implies $\mathcal{N}_{\zeta(\sigma_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath}) \subseteq \mathcal{N}_{h(\sigma_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath})$ for all cases $\langle x_\imath, \lambda_{x_\imath}\rangle$.

Yet, assuming the profile ζ to be unknown, one cannot guarantee the admissibility of a hypothesis h and, hence, the correctness of $C^{est}(x_0)$. In other words, it might happen that $\lambda_{x_0} \notin C^{est}(x_0)$. Fortunately, our results below will show that, using suitable hypotheses, the probability of incorrect predictions is bounded and becomes (arbitrarily) small for large memories.

As will become clear below, a convenient representation of a hypothesis is a step function

$$h : x \mapsto \sum_{k=1}^{m} \beta_k \cdot \mathbb{I}_{A_k}(x), \tag{4}$$

where $A_k = [\alpha_{k-1}, \alpha_k)$ for $1 \leq k \leq m-1$, $A_m = [\alpha_{m-1}, \alpha_m]$, and $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$ defines a partition of $[0,1]$. The strongest hypothesis $h_{\mathcal{D}}$ consistent with the data $\mathcal{D}$ is characterized by the coefficients

$$\beta_k =_{\text{def}} \min_{x_\imath, x_j : \sigma_{\mathcal{X}}(x_\imath,x_j) \in A_k} \sigma_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_j}) \tag{5}$$

for $1 \leq k \leq m$, where $\min \emptyset = 1$ by definition. We call $h_{\mathcal{D}}$ the *empirical similarity profile* and, for obvious reasons, the step function h^* defined by the values

$$\beta_k^* =_{\text{def}} \inf\{\, \zeta(x) \,|\, x \in A_k \,\}, \tag{6}$$

$1 \leq k \leq m$, the *optimal admissible* hypothesis. Since admissibility implies consistency, we have $h^* \leq h_{\mathcal{D}}$. This inequality suggests that the empirical similarity profile $h_{\mathcal{D}}$ will usually overestimate the true profile ζ and, hence, that $h_{\mathcal{D}}$ might not be admissible. Of course, the fact that admissibility of $h_{\mathcal{D}}$ is not guaranteed seems to conflict with the objective of providing correct predictions and, hence, gives rise to questions concerning the actual quality of the empirical profile as well as the quality of predictions derived from that hypothesis. The following theorem gives an important answer to this question.

Theorem 1 *Suppose that observed instances are independent and identically distributed (iid) random variables, generated according to a fixed (not necessarily known) probability distribution μ over $\mathcal{X}$. Let $C^{est}(x_0)$ be the prediction of the label λ_{x_0} derived from the hypothesis $h_{\mathcal{D}}$. The following estimation holds true:*

$$\mathsf{Pr}\left(\lambda_{x_0} \notin C^{est}(x_0)\right) \leq 2m/(1+|\mathcal{D}|), \tag{7}$$

where m is the size of the partition underlying the step function $h_{\mathcal{D}}$. □

According to this result, the probability of an incorrect prediction becomes small for large memories, even if the related hypotheses are not admissible. In fact, $\mathsf{Pr}(\lambda_{x_0} \notin C^{est}(x_0)) \to 0$ as $|\mathcal{D}| \to \infty$. In a statistical sense, the predictions $C^{est}(x_0)$ can indeed be seen as *credible sets*, a justification for using this term not only for $C(x_0)$ but also for $C^{est}(x_0)$. Note that the level of confidence guaranteed by $C^{est}(x_0)$ depends on the number of observed cases and can hence be controlled.

It is furthermore interesting to note that the level of confidence does *not* depend on $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$, i.e. credible predictions can be made for *any* pair of similarity measures. Needless to say, however, the more suitably these measures are chosen, the more *precise* predictions will be.

Let us finally mention that a result similar to the above theorem can also be obtained for the case of *local* similarity profiles [7]. In this case, predictions are usually more precise but less confident (cf. Section 5).

3.3 Practical Issues

The above results provide a sound theoretical basis for an instance-based prediction of credible label sets. In the remainder of this section, we shall discuss some modifications the purpose of which is to improve the practical usefulness of the method.

A rather obvious idea in connection with the inference scheme (3) is to take the intersection not over all cases in $\mathcal{D}$ but only over the $k \ll n$ nearest neighbors of the query x_0. Obviously, this will increase efficiency while preserving the correctness of the prediction. On the other hand, some precision will also be lost, but this effect is usually limited due to the fact that less similar instances often

hardly contribute to the precision of predictions. More specifically, we have implemented the following strategy: The k nearest neighbors are rank ordered according to their similarity to the query, and the intersection (3) is derived in this order successively. Moreover, the prediction of the $\imath$-th nearest neighbor is ignored if its intersection with the current result would produce an empty set (note that this does again preserve correctness).

In many applications one is interested in both, a point-estimation (of a numeric attribute) and a credible set. In this case, the former can of course be derived from the latter, for example as a kind of "center of gravity". In particular, if the credible set takes the form of an interval (as in our experiments in Section 5), an obvious candidate is the mid-point of the interval.

Let us finally make a note on the specification of the similarity measures $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$. Usually, the definition of the latter is uncritical, especially since only the ordinal structure of this measure is important: A (strictly) monotone transformation of $\sigma_{\mathcal{L}}$ will not change the inference results (it changes the similarity bounds $\beta = h(\alpha)$ but not the neighborhoods $\mathcal{N}_\beta(\lambda)$). Particularly, this means that $\sigma_{\mathcal{L}}$ can simply be defined by e.g. a linear function $(\lambda, \lambda') \mapsto 1 - |\lambda - \lambda'|/(u-l)$ for numeric attributes with a bounded range $[l, u]$. As concerns the definition of $\sigma_{\mathcal{X}}$, the cardinal structure is important in so far as it has an influence on the assignment of similarity pairs to the bins of the (fixed) partition underlying the specification of the similarity profile. Still, our experiments have shown the profile to be rather robust toward variations of $\sigma_{\mathcal{X}}$. This can be explained by the fact that moving a similarity pair from one bin to another does only have an effect if this pair is a "critical" one that determines the similarity bound in one of the bins. In practice, we have achieved good results with measures of the form

$$\sigma_{\mathcal{X}}(x, y) \;=_{\text{def}}\; \exp(-\gamma \, \|x - y\|_2). \tag{8}$$

The constant γ is a degree of freedom that must be adapted to the application at hand, e.g. by means of a cross validation. To guarantee that each attribute does approximately have the same influence – a point of critical importance in IBL [9] – the data is first re-scaled linearly to the unit (hyper)cube.

4 PROBABILISTIC PROFILES

A similarity profile ζ defined according to (2) is obviously quite sensitive toward outliers, i.e., similarity pairs

$$(\alpha, \beta) = \big(\sigma_{\mathcal{X}}(x_\imath, x_\jmath), \sigma_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_\jmath}) \big) \tag{9}$$

with small β. In fact, $\zeta(\alpha)$ is a *lower bound* to the similarity of labels that belong to α-similar instances. Thus, even the existence of a single pair of α-similar instances having rather dissimilar labels entails a small lower bound $\zeta(\alpha)$. Small bounds in turn will obviously have a negative effect on the precision of predictions (3). This problem is diminished to some extent by the use of local profiles, since such profiles are derived from a much smaller number of similarity pairs (9). Nevertheless, local profiles are still lower bounds and, hence, not robust toward outliers.

A reasonable idea in this connection is to replace deterministic similarity bounds $\zeta(\alpha)$ by *probabilistic* bounds, that is by (cumulative) probability distribution functions F_α, with $F_\alpha(\beta)$ being the probability that $\sigma_{\mathcal{L}}(\lambda_x, \lambda_y) \leq \beta$ for α-similar instances $x, y \in \mathcal{X}$. In practice it will usually be sufficient to characterize a distribution function by a finite set of quantiles.

The representation of hypotheses in the form of step functions can easily be extended to the above probabilistic setting. Let A_k be an interval in the representation (4) of hypotheses. Moreover, let S_k be the set of similarity degrees $\sigma_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_\jmath})$ such that $\sigma_{\mathcal{X}}(x_\imath, x_\jmath) \in A_k$. Rather than assigning to β_k the minimum of S_k, as in (5), we now define this bound by the $(1-p)$-quantile of S_k, where p is a usually small value such as 0.05. As an empirical quantile, β_k is hence an estimation of the corresponding true quantile of F_α. We call the step function h^p defined by $h^p(\alpha) = \beta_k$ for $\alpha \in A_k$ the *empirical p-profile*.

Now, suppose that we employ h^p in order to derive a prediction

$$C(x_0) \;=\; \bigcap_{\imath=1}^{k} \mathcal{N}_{h^p(\sigma_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath}),$$

where $x_1 \ldots x_k$ are the query's k nearest neighbors. What is the level of confidence of this prediction? Unfortunately, we do not have enough information to compute the probability of an incorrect prediction exactly. Still, by making a simplifying independence assumption à la naive Bayes, the confidence level $(1-p)^k$ can be justified. Our practical experience has shown that this level still underestimates the true confidence level in almost any application (cf. Section 5).

Of course, probabilistic estimations of the above type can be derived for different values $p_1 < p_2 < \ldots < p_\ell$. Thus, one obtains a nested sequence

$$C^{p_\ell}(x_0) \;\subseteq\; C^{p_{\ell-1}}(x_0) \subseteq \ldots \subseteq\; C^{p_1}(x_0)$$

of credible label sets with associated confidence levels. As an advantage of this kind of "stratified" prediction note that it differentiates between predicted labels better than a single credible label set does: The labels in $C^{p_\ell}(x_0)$ are the *most likely* ones, those in $C^{p_{\ell-1}}(x_0) \setminus C^{p_\ell}(x_0)$ are somewhat less likely, and so on.

5 EXPERIMENTAL RESULTS

This section is meant to convey a first idea of the practical performance of CIBL, without laying claim to providing an exhaustive experimental evaluation. In the experiments presented here, we compared our approach to standard IBL (nearest neighbor estimation [5]) and linear regression. We refrained from "tuning" the different methods. Particularly, for IBL we neither included feature selection nor feature weighting. (It is well-known that irrelevant features can badly deteriorate IBL and, on the other hand, that feature weighting can greatly improve performance [15].)

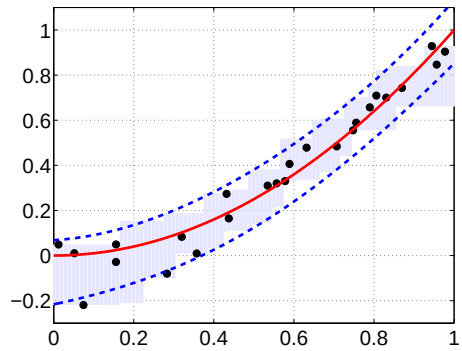

Figure 1. Approximation of $x \mapsto x^2$ (solid line) in the form of a confidence band, using CIBL (shaded region) and linear regression (region between dashed lines). The sample is indicated by black points.

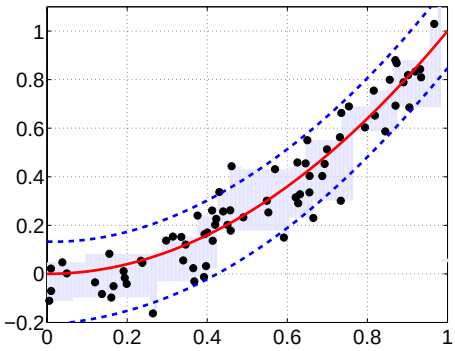

Figure 2. Instance-based approximation using local similarity profiles.

5.1 Artificial Data

The first example is a simple regression problem and mainly serves as an illustration purpose. The function to be learned is given by the polynomial $x \mapsto x^2$. Moreover, n training examples $\langle x_i, \lambda_{x_i} \rangle$ are given, where the x_i are uniformly distributed in $\mathcal{X} = [0, 1]$ and the λ_{x_i} are normally distributed with mean $(x_i)^2$ and standard deviation $1/10$. As a similarity measure for instances, we employed (8) with $\gamma = 2$. Given a random sample $\mathcal{D}$, we first induce a similarity hypothesis for an underlying equi-width partition of size $m = 5$. Using this hypothesis and the sample $\mathcal{D}$, we derive a prediction λ_x for all instances $x \in [0, 1]$ (resp. for the discretization $\{0, 0.01, 0.02 \ldots 1\}$). Note that such a prediction is simply an interval. The union of these intervals yields a *confidence band* for the true mapping $x \mapsto x^2$. Fig. 1 shows a typical inference result for $n = 25$. Moreover, Fig. 2 shows a result for $n = 75$, using local similarity profiles (CIBL-L).

According to our estimation (7), the degree of confidence for $n = 25$ is $16/26$. This, however, is only a lower bound, and empirically (namely by averaging over 1,000 experiments) we found that the level of confidence is almost 0.9. To draw a comparison with standard statistical techniques, the figures also show the 0.9-confidence band obtained for the regression estimation (and the same samples). As can be seen, CIBL yields predictions of roughly the same precision, CIBL-L is even slightly more precise. This finding was also confirmed for estimation problems with a higher dimensional input space, which are not presented here due to reasons of space.

In this connection it deserves mentioning that linear resp. polynomial regression makes much more assumptions than CIBL. Especially, the type of function to be estimated must be specified in advance: Knowing that this function is a polynomial of degree 2 in our example, we took the model $x \mapsto \beta_0 + \beta_1 x + \beta_2 x^2$ as a point of departure and estimated the coefficients β_i, however usually such knowledge will not be available. For instance, typical overfitting effects can be observed when adapting a polynomial of degree $k > 3$ to the data. Moreover, the confidence band is only valid if the error terms follow a normal distribution (as they do in our case but not in general).

5.2 Real-World Data

We also applied our method to several real-world data sets from the UCI repository. Due to reasons of space, we restrict our discussion to results for the `auto-mpg` data. This data set contains the city-cycle fuel consumption in miles per gallon for 392 cars (with values between 9 and 46.6), to be predicted in terms of 3 multivalued discrete and 4 continuous attributes. In order to facilitate the comparison with linear regression, we only used the 4 continuous attributes (displacement, horsepower, weight, acceleration).

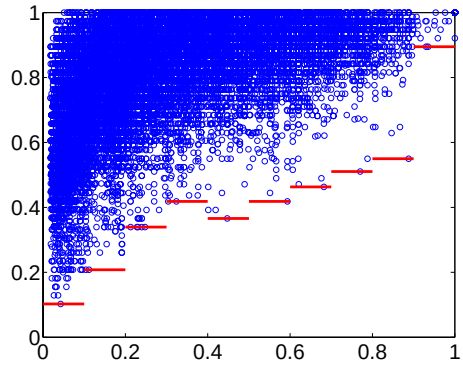

Figure 3. Similarity profile for the `auto-mpg` data (step function). Each point corresponds to a pair (α, β) with $\alpha = \sigma_{\mathcal{X}}(x, y)$ (abscissa) and $\beta = \sigma_{\mathcal{L}}(\lambda_x, \lambda_y)$ (ordinate).

Fig. 3 shows the similarity profile for the data, using a partition of size 10. The picture clearly reveals the aforementioned outlier effect: The similarity profile is "pressed down" by a relatively small number of similarity pairs $(\alpha, \beta) = (\sigma_{\mathcal{X}}(x, y), \sigma_{\mathcal{L}}(\lambda_x, \lambda_y))$. The similarity between instances was measured by (8) with $\gamma = 3$.

In order to test the effectiveness of the probabilistic strategy for CIBL, we have applied this approach to the data with different values for p. The following performance measures were derived by means of a leave-one-out cross-validation: (1) The *precision* of predictions (PREC) measured in terms of the average length of a predicted interval. (2) The *mean absolute error* (MAE) measured in terms of the average distance between the true value and the point estimation (center of the interval). (3) The *correctness* or empirical confidence (CONF) measured in terms of the relative frequency of correct predictions (predicted interval covers true value). The following table shows results for different sizes k of the neighborhood:

k	p	CONF	PREC	MAE
3	.00	1.00	37.52	2.90
3	.02	0.96	23.21	2.92
3	.04	0.93	19.65	2.93
7	.00	1.00	34.84	2.92
7	.02	0.92	19.85	2.93
7	.04	0.88	16.11	2.94
15	.00	1.00	33.00	3.04
15	.02	0.86	17.21	3.13
15	.04	0.77	13.34	3.13

As can be seen, the use of probabilistic bounds yields an extreme gain of precision at the cost of a rather slight deterioration of the MAE. For example, for $k = 15$, the precision is almost doubled when passing from the (deterministic) profile $(p = 0)$ to the empirical p-profile with $p = .02$.

It is also interesting to compare the theoretical confidence level $(1-p)^k$, justified by an assumption of independence, to the empirical confidence level (CONF). Fig. 4 shows the ratio between the latter and the former as a function of p, and for different values k. As can be seen, the ratio is always ≥ 1, i.e. the empirical confidence is always underestimated by the theoretical one, and this underestimation does even increase with p and k.

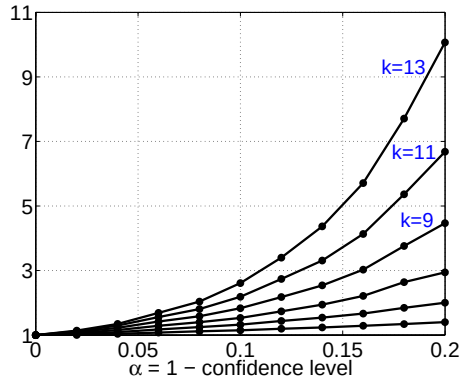

Figure 4. Ratio between theoretical and empirical level of confidence for $\alpha = 0, 0.02 \ldots 0.2$ and $k = 3, 5 \ldots 13$.

We have made the same experiment using the local variant CIBL-L. Since there are less cases involved in deriving the individual profiles, the size of the partition was reduced from 10 to 5. The results, summarized in the table below, confirm our theoretically founded expectations: Predictions become more precise but less confident. Apart from that, it is interesting to note that CIBL-L also yields better point estimations.

k	p	CONF	PREC	MAE
3	.00	0.94	11.81	2.79
3	.02	0.77	8.57	2.72
3	.04	0.77	7.98	2.69
7	.00	0.92	9.98	2.67
7	.02	0.68	7.00	2.61
7	.04	0.66	6.47	2.58
15	.00	0.88	8.99	2.60
15	.02	0.62	6.06	2.55
15	.04	0.60	5.57	2.54

For comparison purposes, we have also repeated the experiment using standard IBL (k-nearest neighbor estimation) and linear regression as prediction methods. IBL predictions were derived as a linear combination of the values from the k nearest neighbors, with the weight of a neighbor being proportional to its similarity. The results are summarized in the following table:

k	3	7	11	15	19
MAE	2.87	2.76	2.81	2.80	2.81

As can be seen, the estimations of CIBL are only slightly worse than those of IBL, and the estimations of CIBL-L are even better. Recalling that CIBL is actually not intended to produce point estimations, this is a surprisingly good result. Apart from that, standard IBL does of course not provide confidence estimations, which is the primary concern and key advantage of CIBL.

For LR, the model fit is acceptable,[2] but the results are not competitive: The mean absolute error is 3.23 and the precision is 19.71 for $p = 0.02$, 16.59 for $p = 0.05$ and 13.91 for $p = 0.1$. Polynomial regression (of orders 2 and 3) yields slightly more accurate point estimations but even less precise confidence intervals.

[2] The R^2-statistic is around 0.71 and the F-statistic is extremely large.

6 CONCLUDING REMARKS

We have proposed an instance-based learning method called CIBL that allows for deriving estimations in the form of *credible label sets*, which are provably correct with high probability (under standard assumptions on the data generating process). Thus, CIBL combines advantages from both, instance-based and model-based (statistical) learning: As an instance-based approach it requires less structural assumptions than (parametric) statistical methods, and yet it allows for specifying the uncertainty related to predictions.

As a further advantage of CIBL let us mention that it hardly assumes more than the specification of similarity measures over instances and labels and, hence, is quite general and universally applicable. Especially, no distinction is actually made between classification and regression. Indeed, CIBL can easily be applied to other types of problems as well, such as e.g. the prediction of *rankings* [6]. In fact, note that no kind of transitivity is assumed for the similarity measures, which means that the structure of $\mathcal{X}$ and $\mathcal{L}$ might be weaker than that of a metric space. Consequently, CIBL is applicable in many situations where standard methods from statistics cannot be used.

A main concern in this paper was the *correctness* of predictions. Let us finally mention that one can also obtain estimations related to the *precision* of predictions. For instance, a result similar to a theorem in [9] can be shown provided that the mapping $x \mapsto \lambda_x$ satisfies certain continuity assumptions (this of course also requires that $\mathcal{X}$ and $\mathcal{L}$ are metric spaces). Namely, this mapping can be approximated to any degree of accuracy, that is, for $\epsilon > 0$ there is a finite memory $\mathcal{D}$ such that $\lambda_x \in C^{est}(x)$ for all $x \in \mathcal{X}$ and $\sup_{x \in \mathcal{X}} \text{diameter}(C^{est}(x)) < \epsilon$.

REFERENCES

[1] D.W. Aha, 'Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms', *International Journal of Man-Machine Studies*, **36**, 267–287, (1992).

[2] *Lazy Learning*, ed., D.W. Aha, Kluwer Academic Publ., 1997.

[3] D.W. Aha, D. Kibler, and M.K. Albert, 'Instance-based learning algorithms', *Machine Learning*, **6**(1), 37–66, (1991).

[4] C.G. Atkeson, A.W. Moore, and S. Schaal, 'Locally weighted learning', *Artificial Intelligence Review*, **11**, 11–73, (1997).

[5] *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed., B.V. Dasarathy, IEEE Comp. Soc. Press, Los Alamitos, 1991.

[6] S. Har-Peled, D. Roth, and D. Zimak, 'Constraint classification: a new approach to multiclass classification', in *Proceedings 13th Int. Conf. on Algorithmic Learning Theory*, pp. 365–379, Lübeck, Germany, (2002).

[7] E. Hüllermeier, 'Focusing search by using problem solving experience', in *Proc.* ECAI–2000, pp. 55–59, Berlin, Germany, (2000).

[8] E. Hüllermeier, 'Instance-based learning of credible label sets', in *Proc.* KI–03, *26th German Conf. on Artificial Intelligence*, Hamburg, (2003).

[9] D. Kibler and D.W. Aha, 'Instance-based prediction of real-valued attributes', *Computational Intelligence*, **5**, 51–57, (1989).

[10] M Kukar and I Kononenko, 'Reliable classifications with machine learning', in *Proc. ECML–02*, pp. 219–231, (2002).

[11] K Proedrou, I Nouretdinov, V Vovk, and A Gammerman, 'Transductive confidence machines for pattern recognition', in *Proc. ECML–02*, pp. 381–390, (2002).

[12] B. Smyth and E. Mc Kenna, 'Building compact competent case-bases', in *Proc. ICCBR–99, 3rd International Conference on Case-Based Reasoning*, pp. 329–342, (1999).

[13] C. Stanfill and D. Waltz, 'Toward memory-based reasoning', *Communications of the ACM*, 1213–1228, (1986).

[14] V.N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[15] D. Wettschereck, D.W. Aha, and T. Mohri, 'A review and empirical comparison of feature weighting methods for a class of lazy learning algorithms', *AI Review*, **11**, 273–314, (1997).